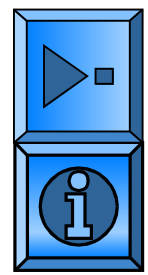
Active Recommendation Project

Collective Organization of the Structure and Sematics of DIS

Luis M. Rocha, Cliff Joslyn, and Johan Bollen
Distributed Knowledge Systems and Modeling Team
Computer Research and Applications Group (CIC-3)
Los Alamos National Laboratory, MS B256
Los Alamos, NM 87545

e-mail: rocha@lanl.gov or rocha@santafe.edu
URL: http://www.c3.lanl.gov/~rocha [/ps/ARP2000.pdf]



DIS and Information Retrieval

- ***Distributed Information Systems*** (DIS): Heterogeneous Collections of electronic networked information resources in interaction with diverse communities of users
  - < the Internet, the World Wide Web, corporate intranets, databases, library information retrieval systems, etc.
- ***Information Retrieval*** (IR): methods and processes for searching relevant information out of information systems (e.g. databases) that contain extremely large numbers of documents
  - < based solely on *keywords* that index (semantically characterize) documents and a query language to retrieve documents from centralized databases in terms of these keywords.

Flaws of Information Retrieval

No User-DIS Structural or Semantic Coupling

- ***Passive Environments***. No genuine interaction between user and system
  - < users *pulls* information from passive database
  - < Users need to know how to query relevant information with appropriate keywords
  - < No user-specific response or recognition (user profiles)
- ***Idle Structure***. Structural relationships between documents, keywords, and information retrieval patterns are not utilized.
  - < e.g. citation structure, WWW link structure, clustering of keyword-document relationships (LSI), temporal patterns of retrieval, etc.
- ***Fixed Semantics***. Keywords do not reflect evolving user semantics
- ***Isolated Information Resources***. No communication among documents and/or keywords in different information resources

Limitations of User-DIS Interaction

- No ***recommendation***. Because of passive environments and idle structure, DIS do not ***recognize*** users and pro-actively recommend appropriate information about related topics that they may be unaware of.
- No ***conversation*** between users and information resources, between information resources, and between users. Because of passive environments and isolated information resources there is no mechanism to exchange and ***crossover*** relevant information.
- No ***creativity*** or ***evolution***. Because of fixed semantics, isolated information resources, idle structure, and passive environments, there is no mechanism to recombine knowledge in different information resources to produce new categories of keywords used by different communities of users.

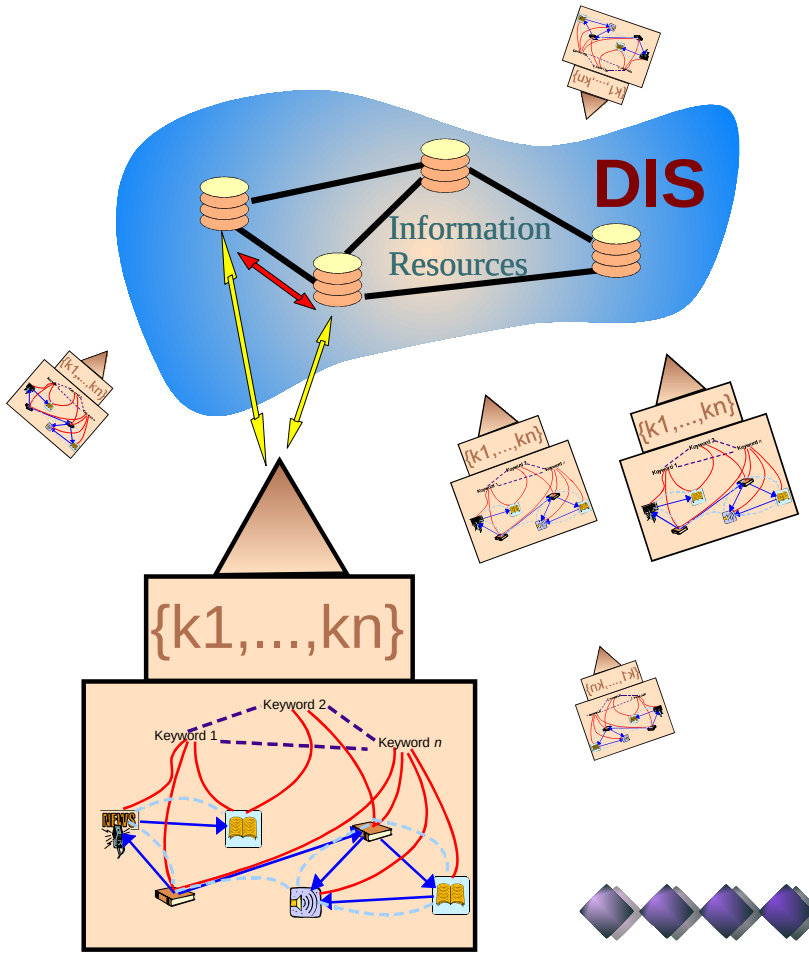
Drawing from Biology

- Distributed Biological Networks
  - < Elicit appropriate responses to sepecific demands
  - < Process relevant information across the network
  - < Adapt to a changing environment by creating novel behaviors
- Computers possess full programmability but no inherent evolvability (Conrad)
  - < *Enabling substrate* (chemistry) for dynamic agent-environment coupling
- Computer systems need enabling relationship-packages
  - < Mechanisms for *active* structural and semantic coupling between agents and distributed systems: *collective organization*
  - < *Adaptability*: capacity to change with feedback
  - < *Evolvability*: Building blocks that can be re-combined, to produce new behavior or function that is not fully pre-specified

Active Recommendation Systems

Distributed Structural and Semantic Coupling

- A means to recognize *users* (agents)
- A means to characterize *information resources*
- *Conversation* mechanisms between users and information resources
- *Adaptation* mechanisms

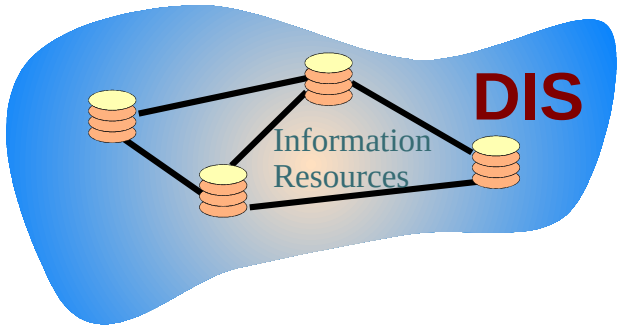


http://www.c3.lanl.gov/~rocha/lww

Active Recommendation Systems for The Library Without Walls

Information Resources:Distributed Memory

The XML Repository



XML Records

- ARPID
- Citation
- Biblio
- Keyterms
- Code
- Other info

XML Repository

Information Resources

Database 1

- Document Records

Database 2

- Document Records

Actual Documents

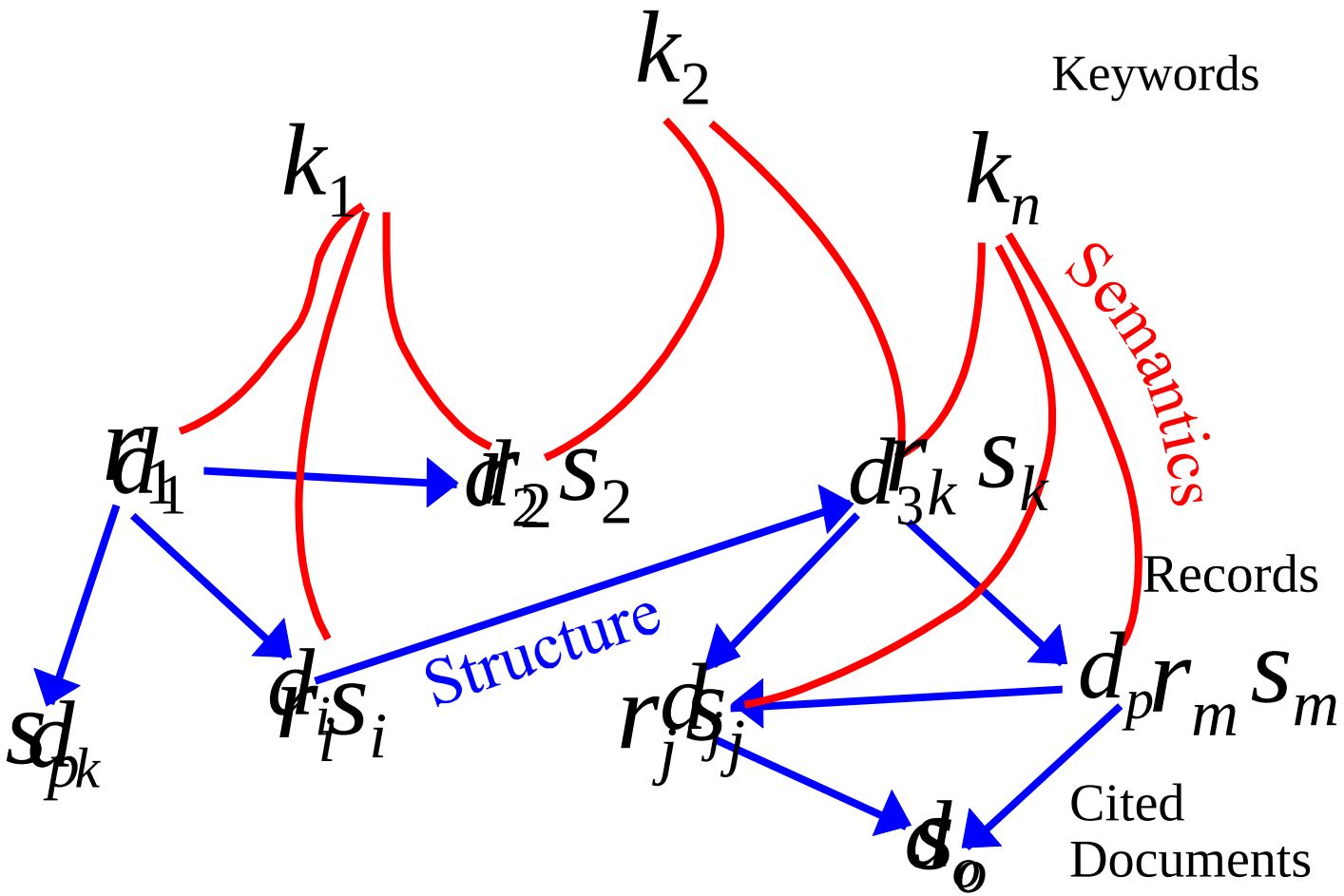
XMLRepository

ExampleofaRecord

```
<RECORD>
        <ARPID>ISSN_1013-9826_1998_137_55</ARPID>
        <CITATION>
            <REFID>ISSN_0032-3861_1987_28_1489 <~>
            ISSN_0022-2461_1994_29_3377 <~>ISSN_0032-3861_1994_35_3948 <~>
            ISSN_0032-3861_1995_36_4587 <~> ISSN_0032-3861_1995_36_4605 <~>
            ISSN_0032-3861_1995_36_4621 <~> ISSN_0022-2461_1989_24_298 <~>
            ISSN_0032-3861_1980_21_466 <~> ISSN_0032-3861_1985_26_1855
          </REFID>
        </CITATION>
        <BIBLIO>
            <TITLE>Effect of rubber functionality on mechanical and fracture properties of impact-modified
nylon 6,6/polypropylene blends</TITLE>
            <ENUM TYPE="ENDPAGE">62</ENUM>
        </BIBLIO>
        <KEYTERMS>
            <KEYW TYPE="TITLE">rubber <~> properti <~> nylon <~> mechan <~> impact-modifi <~>
function <~> fractur <~> blend <~> /polypropylen</KEYW>
            <KEYW TYPE="KEYW_AU">PA6,6/PP blends <~> rubber-toughened nylon <~> rubber-
toughened polypropylene <~> mechanical properties <~> fracture toughness <~> J(c) <~>
fractography</KEYW>
            <KEYW TYPE="KEYW_ISI">FILLED COMPOSITE-MATERIALS <~> POLYPROPYLENE BLENDS
<~> BLOCK-COPOLYMERS <~> PREDICTIVE MODEL <~> COMPATIBILIZATION <~> POLYAMIDES <~>
MORPHOLOGY <~> CAVITATION <~> PARTICLES</KEYW>
            <KEYW TYPE="AUTHOR">Wong, SC <~> Mai, YW</KEYW>
        </KEYTERMS>
    </RECORD>
```

Information Resources: Distributed Memory

Relational Repository

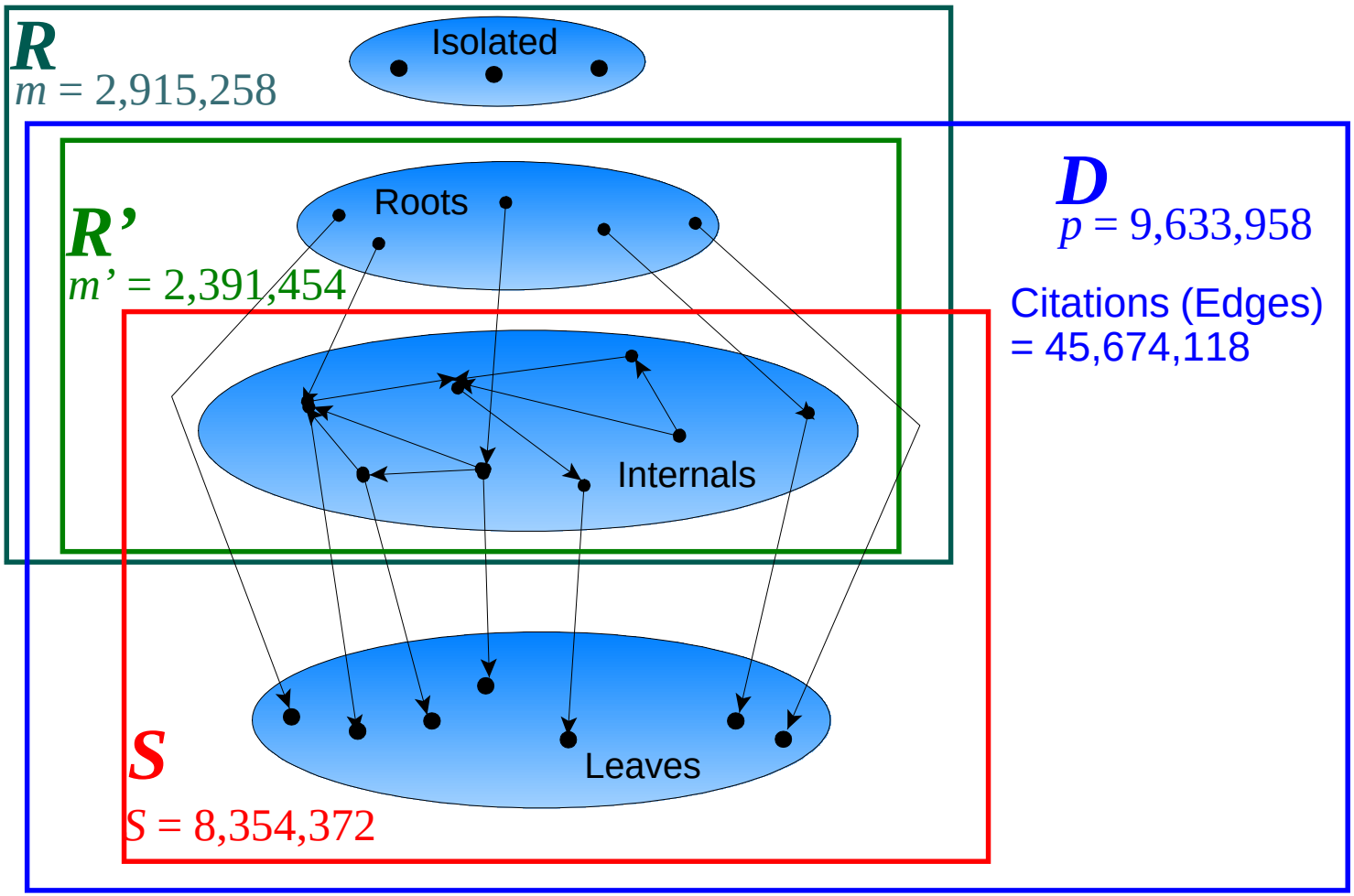


RéS, SéR
D=RcS

Document
Citation Set

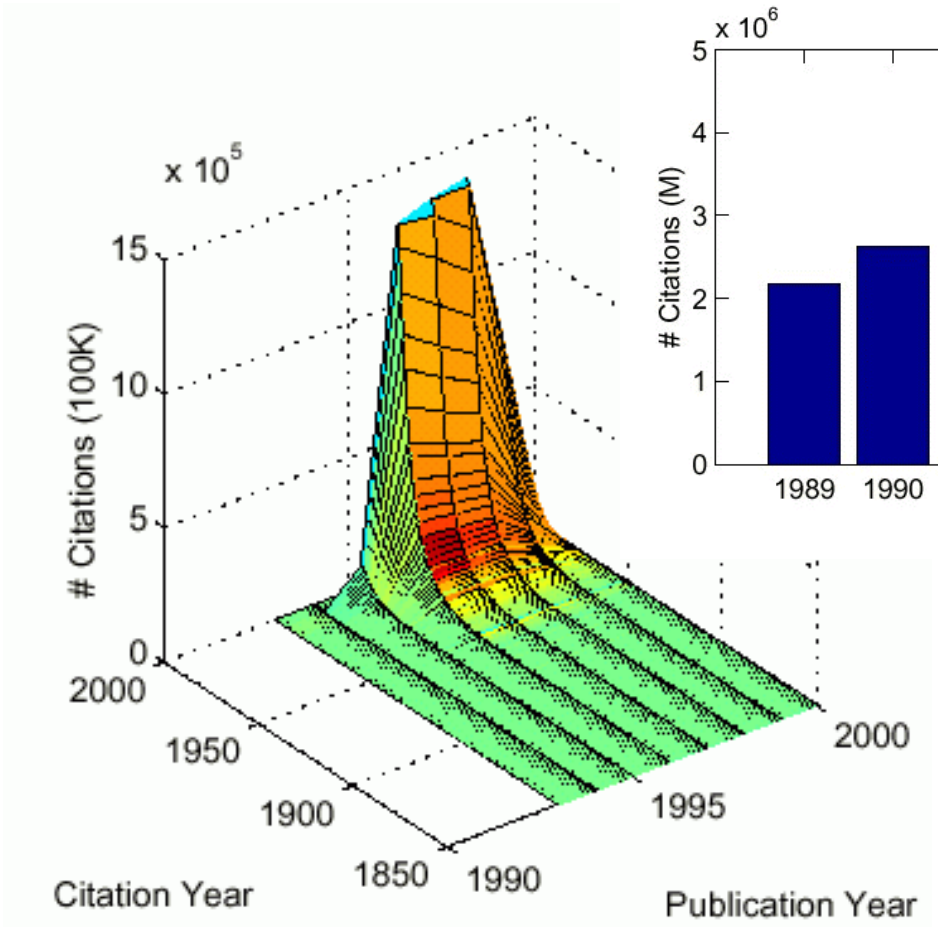
Citation Structure

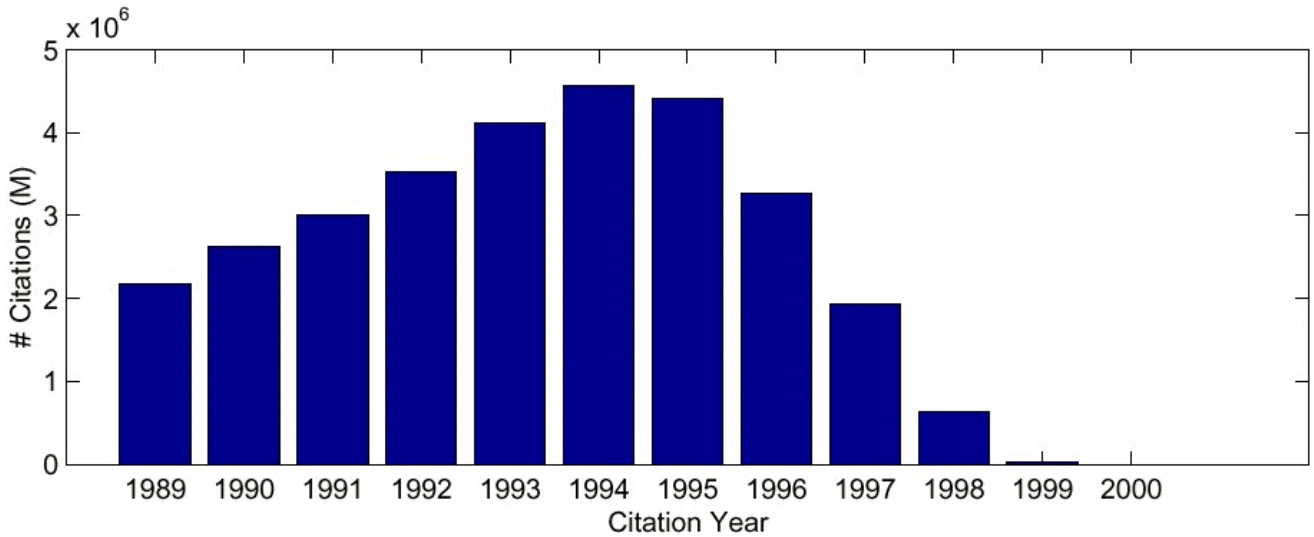
ISNumbers



Citation Analysis

Cited Documents





Citation Analysis

Citation Year	Publication Year						
	1993	1994	1995	1996	1997	1998	1999
1989	25	340	74,803	686,471	614,672	551,898	254,039
1990	21	349	88,453	822,012	737,429	665,398	306,917
1991	21	464	99,925	945,297	850,322	766,659	354,268
1992	1	435	113,803	1,095,707	998,036	900,454	416,636
1993	0	198	127,619	1,261,244	1,167,558	1,061,451	493,942
1994	0	67	114,036	1,337,943	1,319,427	1,222,199	573,001
1995	0	4	40,625	974,363	1,376,726	1,371,412	657,296
1996	0	0	196	185,910	976,039	1,399,689	714,313
1997	0	0	0	371	191,541	1,023,584	726,122
1998	0	0	0	0	541	197,224	444,002
1999	0	0	0	0	0	311	31,812
2000	0	0	0	0	0	0	2

Citation Analysis

Cutoff Statistics

Cutoff Year	Edges	Papers	Citations	Roots	Internals	Leaves	Nodes	Edges/Node
None	45,674,118	2,391,454	8,354,372	1,279,508 0.13	1,111,868 0.12	7,242,504 0.75	9,633,959	4.74
1993	18,990,768	2,121,668	3,075,216	1,086,649 0.26	1,035,019 0.25	2,040,197 0.49	4,161,865	4.56
1994	14,878,756	1,990,772	2,536,261	1,012,568 0.29	978,204 0.28	1,566,077 0.44	3,548,849	4.19
1995	10,312,083	1,722,667	1,943,339	697,014 0.31	835,673 0.30	1,107,665 0.39	2,630,352	3.64
1996	5,891,657	1,231,908	1,329,336	695,316 0.34	536,592 0.27	792,744 0.39	2,024,652	2.91
1997	2,615,510	706,765	757,077	463,719 0.38	243,066 0.20	514,011 0.42	1,220,796	2.14
1998	673,892	266,926	286,761	206,463 0.42	60,463 0.12	226,298 0.46	493,224	1.37
1999	32,125	22,526	24,365	18,979 0.44	3,547 0.08	20,818 0.48	43,344	0.74

Structure in Relational Repository

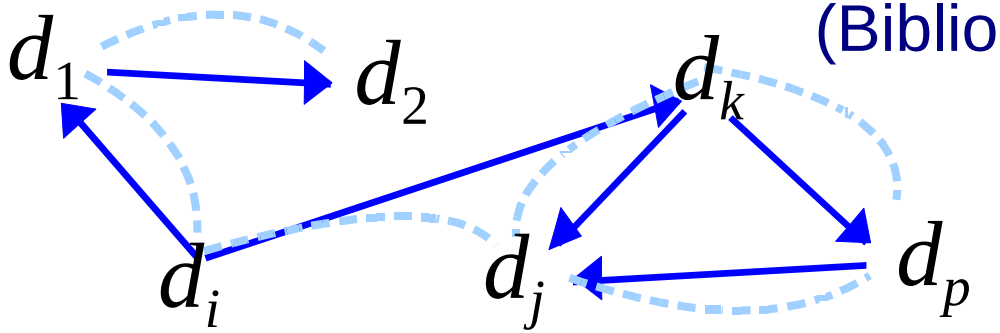
Proximity Relations

$$p^{\text{in}}(d_i, d_j) = \frac{\sum_{k=1}^{p} (c_{k,i} \wedge c_{k,j})}{\sum_{k=1}^{p} (c_{k,i} \vee c_{k,j})} = \frac{N_{\cap}^{\text{in}}(d_i, d_j)}{N_{\cup}^{\text{in}}(d_i, d_j)} = \frac{N_{\cap}^{\text{in}}(d_i, d_j)}{N^{\text{in}}(d_i) + N^{\text{in}}(d_j) - N_{\cap}^{\text{in}}(d_i, d_j)}$$

Inwards Structural Proximity (co-citation)

$$p^{\text{out}}(d_i, d_j) = \frac{\sum_{k=1}^{p} (c_{i,k} \wedge c_{j,k})}{\sum_{k=1}^{p} (c_{i,k} \vee c_{j,k})} = \frac{N_{\cap}^{\text{out}}(d_i, d_j)}{N_{\cup}^{\text{out}}(d_i, d_j)} = \frac{N_{\cap}^{\text{out}}(d_i, d_j)}{N^{\text{out}}(d_i) + N^{\text{out}}(d_j) - N_{\cap}^{\text{out}}(d_i, d_j)}$$

Outwards Structural Proximity (Bibliographic Coupling)



Structural Proximity (linear combination of outwards and inwards) $p_{i,j}$

Documents

- Defines a neighborhood semi-metric (set of documents related to document d_i with proximity greater than " 0 [0, 1])

Structure Analysis

- Clusters of Related Documents
  - Helps with local distribution of tasks
- Kleinberg's Authoritative Sources
  - Reduces the number of relevant documents
- Watts' Small-World Graphs
  - What happens with adaptation?

Keyword Semantics

Tokens of User Knowledge in Communities

- Keywords n=839,297
  - < Author Supplied
  - < Editor Supplied
  - < Titles
- Citations

ARPA Records (ISI 96-99)
m = 2,915,258

Information Resources

Documents

187705	cell
150795	studi
149594	system
140738	express
127350	protein
124094	model
120215	activ
113740	human
112737	rat
112702	patient
104850	effect
102934	gene
92265	structur
90620	analysi
88954	Growth
80818	receptor
79324	mechan
72125	acid
71201	complex
66444	cancer
64091	function

Author Supplied

7834	rat
7405	apoptosis
6081	nitric oxide
5741	immunohistochemistry
4906	epidemiology
4009	crystal structure
3976	development
3912	hypertension
3825	children
3742	cytokines
3709	human

Title

152808	effect
130160	cell
119457	studi
87477	patient
83529	human
80558	system
78915	activ
74003	structur
70344	analysi
66665	new
65767	rat
65675	protein

Editor Supplied

9866	POLYMERASE CHAIN-REACTION
9854	PATTERNS
9750	FLOW
9691	MONOCLONAL-ANTIBODIES
9620	POPULATION
9515	PREVALENCE
9511	SACCHAROMYCES-CEREVISIAE
9431	SCATTERING
9379	ASSOCIATION
9211	STIMULATION
9083	TUMORS
9057	ADSORPTION
9031	RECEPTORS
8953	AGE

Semantics in Relational Repository

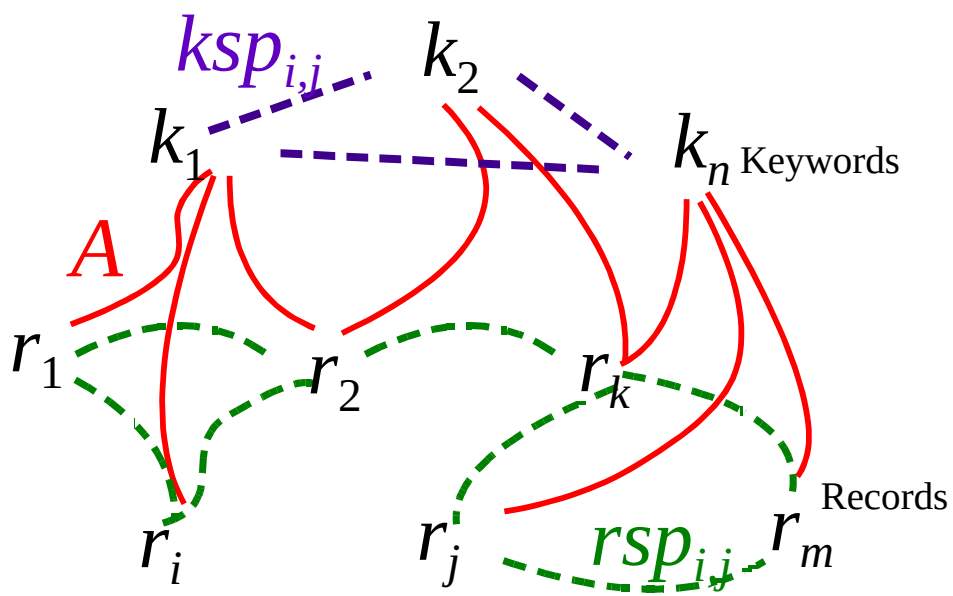
Proximity Relations

$$ksp(k_i, k_j) = \frac{\sum_{k=1}^{m}(a_{i,k} \wedge a_{j,k})}{\sum_{k=1}^{m}(a_{i,k} \vee a_{j,k})} = \frac{N_{\cap}(k_i, k_j)}{N_{\cup}(k_i, k_j)}$$

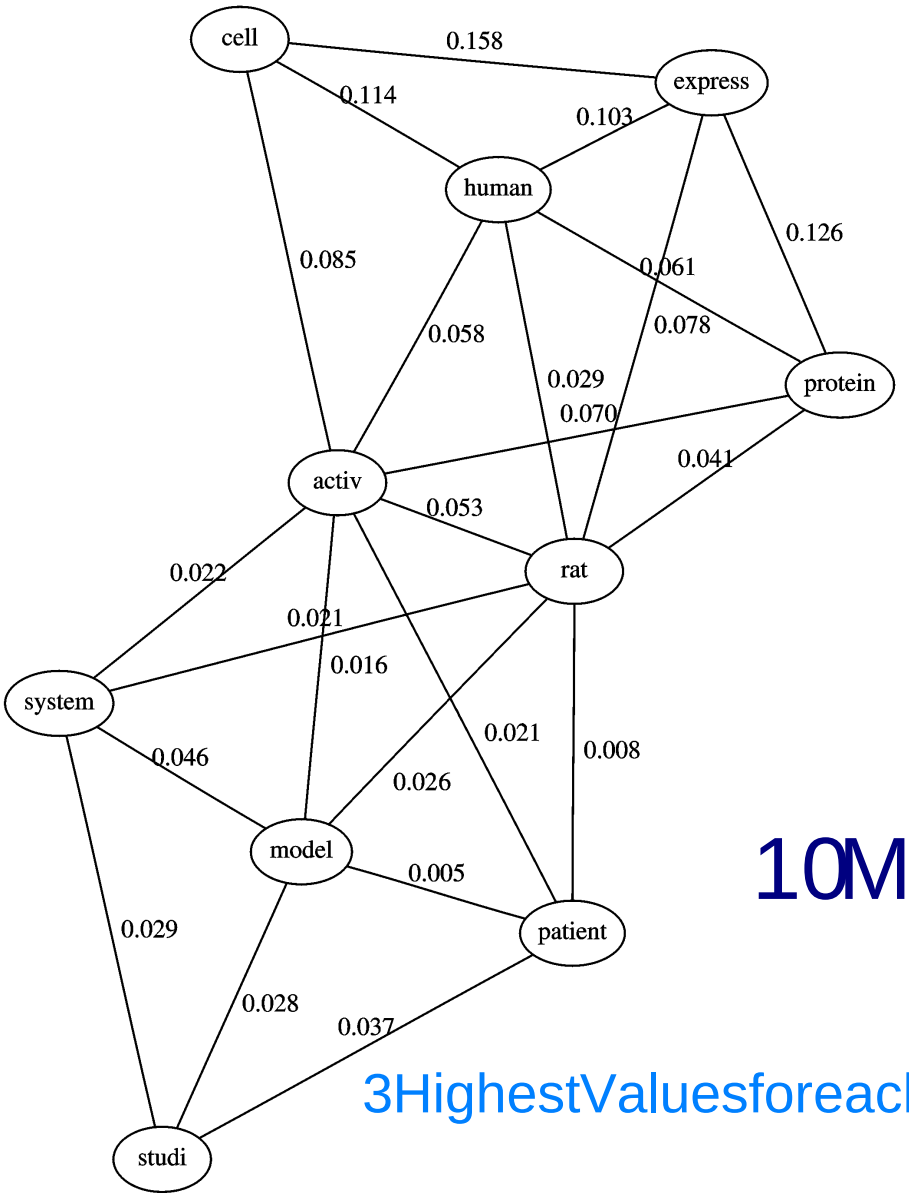
(Keyword Semantic Proximity)

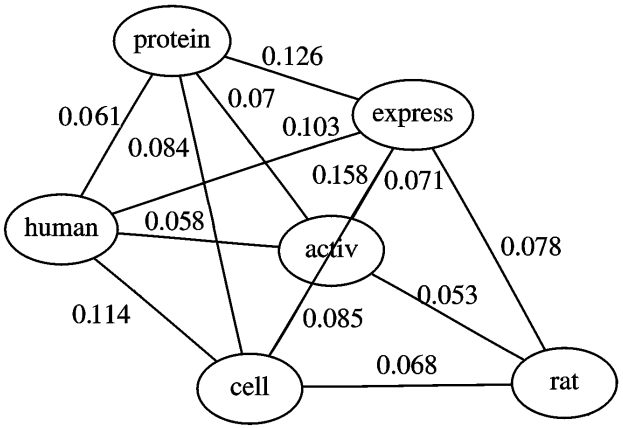
$$rsp(r_i, r_j) = \frac{\sum_{k=1}^{m}(a_{k,i} \wedge a_{k,j})}{\sum_{k=1}^{m}(a_{k,i} \vee a_{k,j})} = \frac{N_{\cap}(r_i, r_j)}{N_{\cup}(r_i, r_j)}$$

(Record Semantic Proximity)

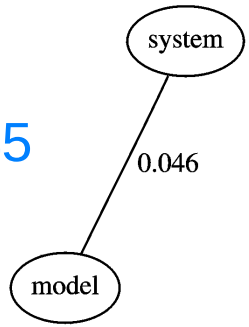


Keywords Semantic Proximity





Values Higher than 0.045



10 Most Common Keywords

3 Highest Values for each node

Keyword Semantic Proximity

10 Most Frequent Keywords

	cell	studi	system	express	protein	model	activ	human	rat	patient
cell	1.000	0.022	0.019	0.158	0.084	0.017	0.085	0.114	0.068	0.032
studi	0.022	1.000	0.029	0.013	0.017	0.028	0.020	0.020	0.020	0.037
system	0.019	0.029	1.000	0.020	0.017	0.046	0.022	0.014	0.021	0.014
express	0.158	0.013	0.020	1.000	0.126	0.011	0.071	0.103	0.078	0.020
protein	0.084	0.017	0.017	0.126	1.000	0.013	0.070	0.061	0.041	0.014
model	0.017	0.028	0.046	0.011	0.013	1.000	0.016	0.016	0.026	0.005
activ	0.085	0.020	0.022	0.071	0.070	0.016	1.000	0.058	0.053	0.021
human	0.114	0.020	0.014	0.103	0.061	0.016	0.058	1.000	0.029	0.021
rat	0.068	0.020	0.021	0.078	0.041	0.026	0.053	0.029	1.000	0.008
patient	0.032	0.037	0.014	0.020	0.014	0.005	0.021	0.021	0.008	1.000

http://www.c3.lanl.gov/~rocha/lww/keywords.html

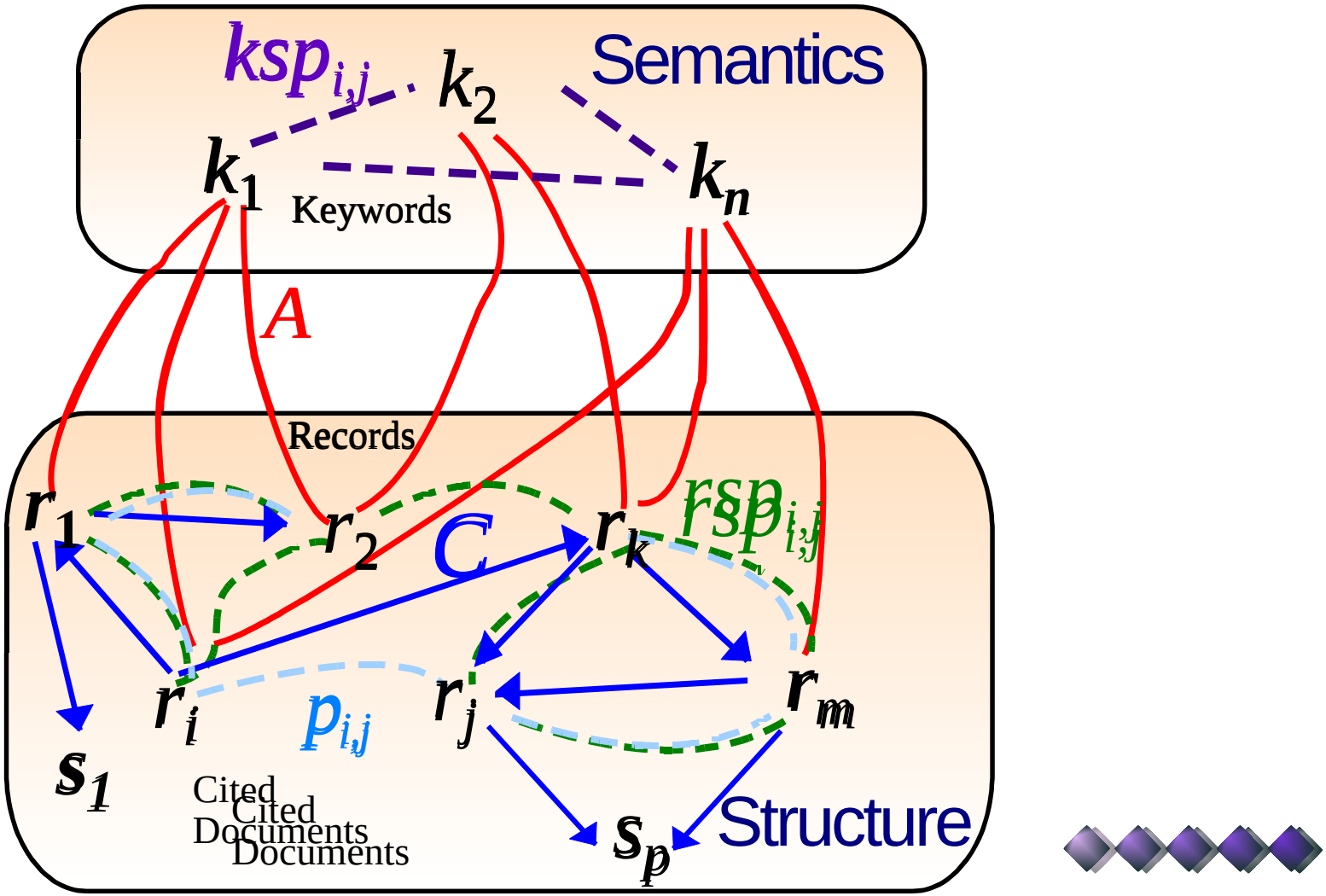
- Most Frequent Keyword Proximities Calculated
- Other Proximities calculated as needed

Semantic Analysis

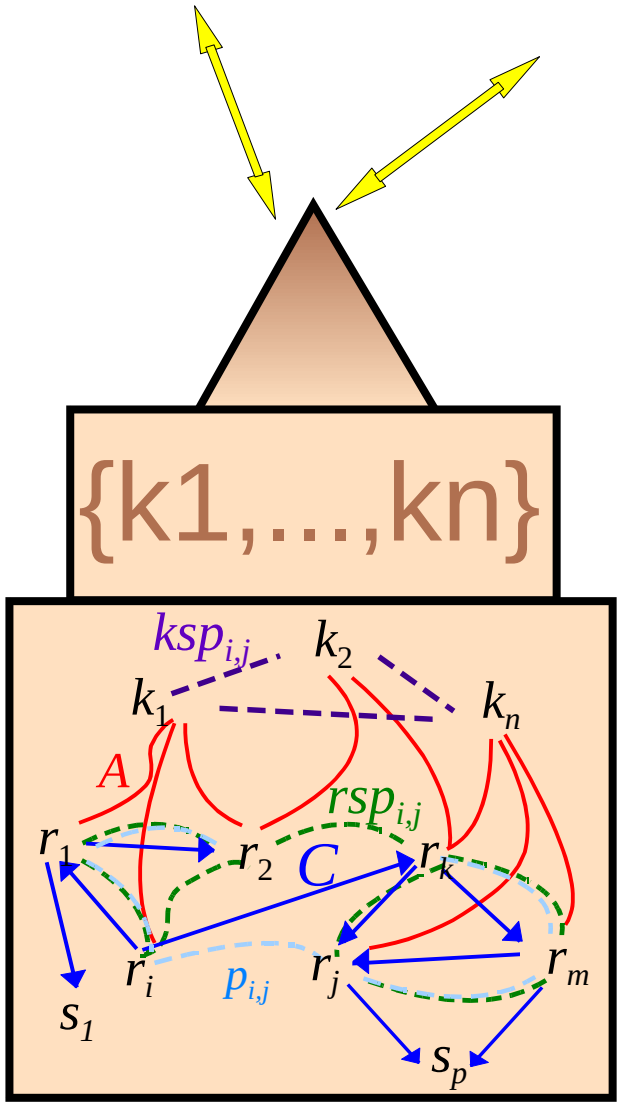
- Clusters of Semantically Related Documents
  - Helps with local distribution and classification
- Latent Semantic Indexing
  - Reduces the number of relevant documents and Keywords
- Metricity
  - How does it change with adaptation?
- Small-world graphs

KnowledgeContext

InformationResourceforaCommunityofUsers



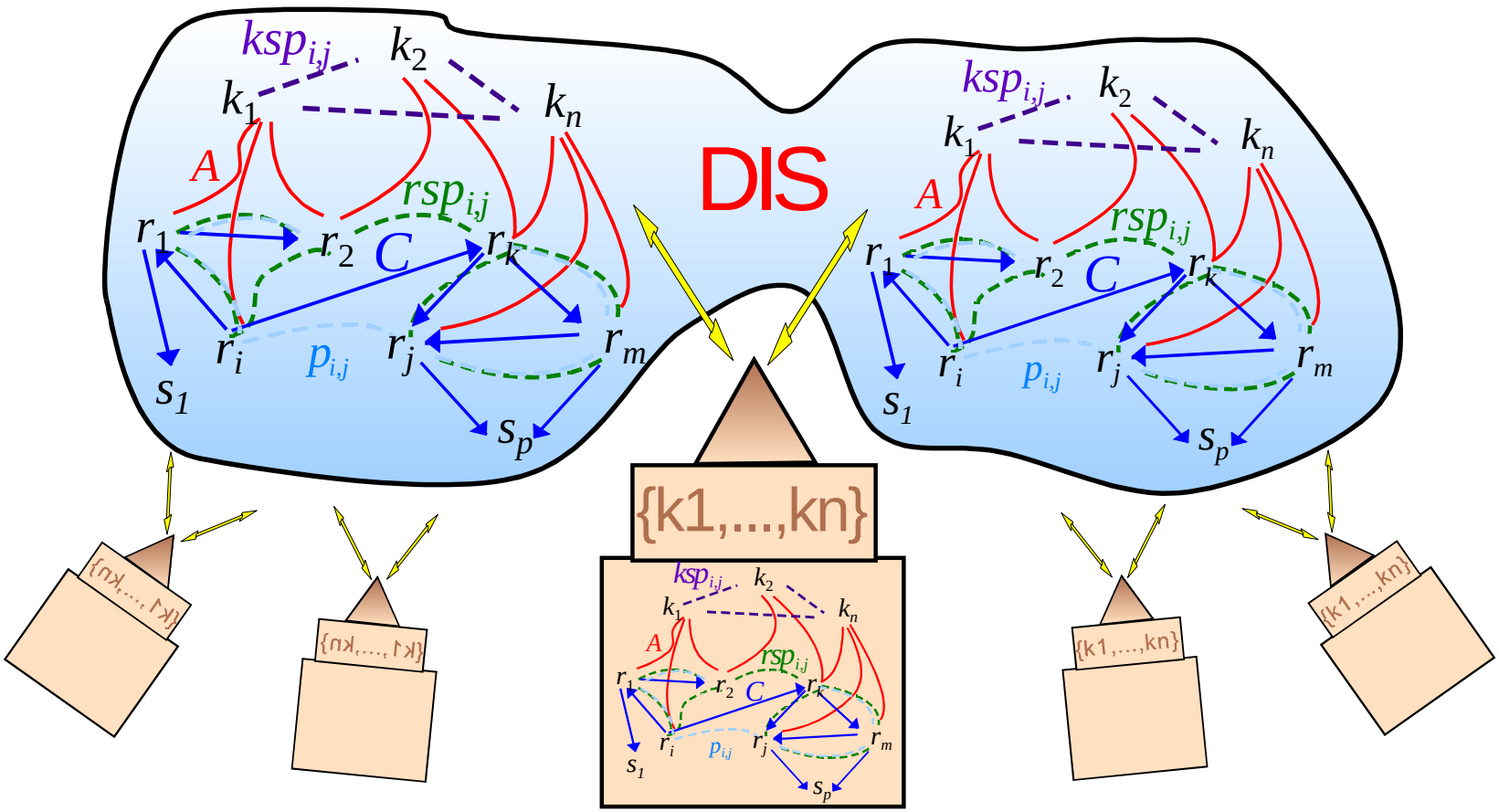
Users



- ***Present interests***: $\{k_1, \ldots, k_i\}$.
- ***History of IR***: a knowledge context.
  - < users as information resources with their own proximity information.
- ***Communication Protocol***. 2-way means to communicate with other information resources
  - < Retrieval, adaptation

DIS and Users

Collective Interaction



- Enabling Substrate for Active, Biologically Motivated, Recommendation
  - < User Interests Can now be Recognized and Related to Stored Knowledge

Communication

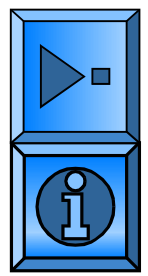
Categorization, Linguistic Recombination

We need a mechanism to enable the communication between users/agents and information resources, leading to information exchange, adaptation and recombination.

A Model of (Clark's) "on the hoof" Categories

Evidence Sets

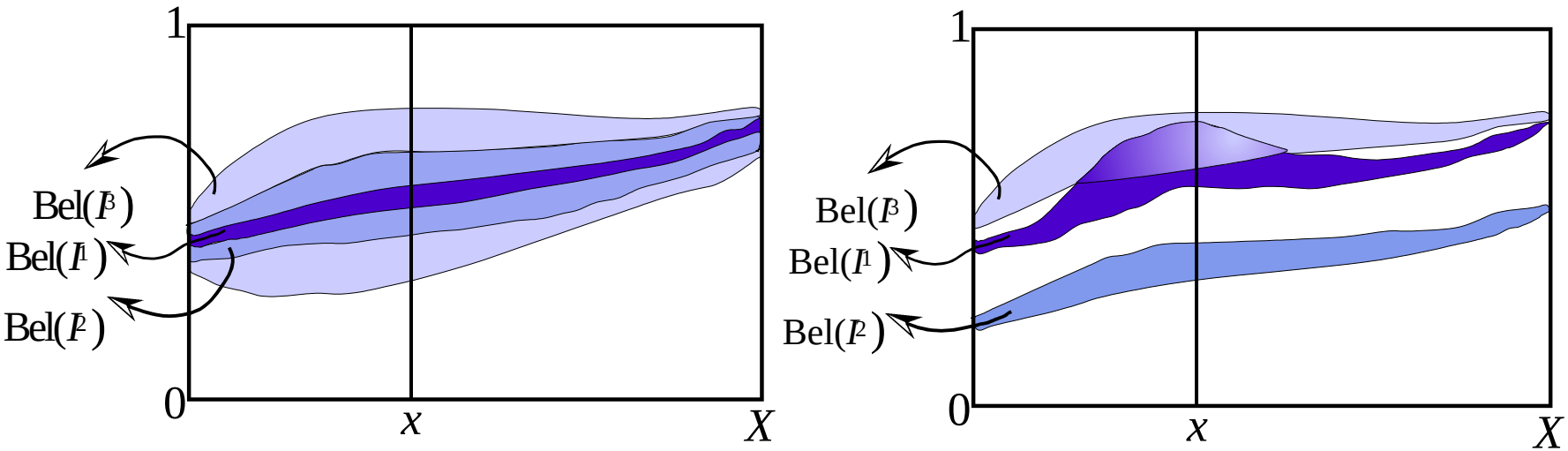
$$A(x): X \rightarrow \mathbf{B}[0,1]$$



[http://www.c3.lanl.gov/~rocha/ijhms_pask.html; dissert.html; ijgs_unc.html; es_ijgs.html; kluwer.html]

Evidence Sets

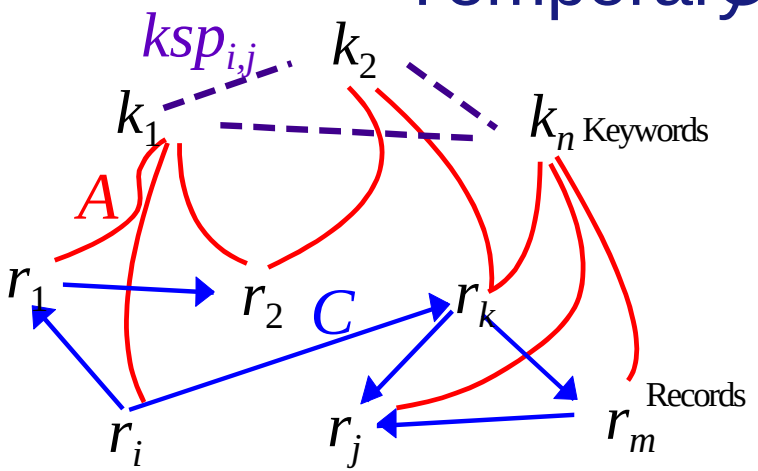
Evidentially Graded Interval Membership



- Set-theoretic (or logic) framework
  - < Complement, Union, Intersection, etc...
- Measures of Uncertainty
  - < Nonspecificity, Conflict, Fuzziness

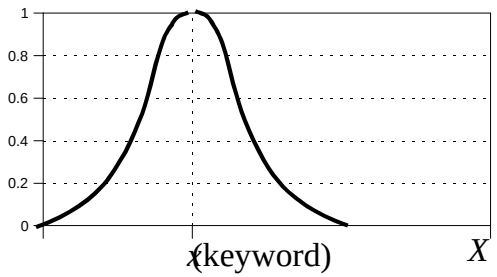
"On the hoof" Categories

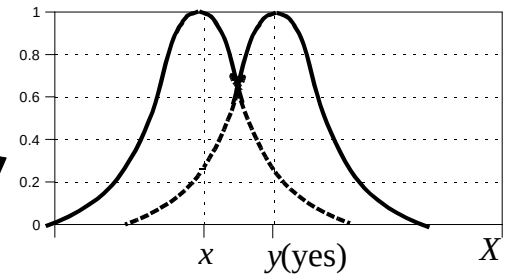
Temporary Communication Constructs

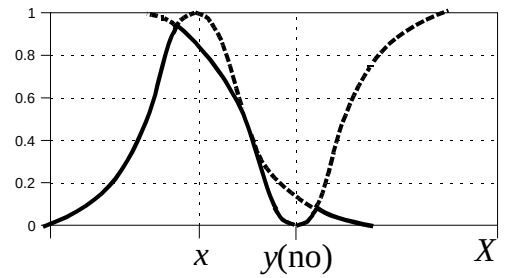


No Categories are stored as such, but rather constructed in conversation between agents and information resources, leading to agent identification

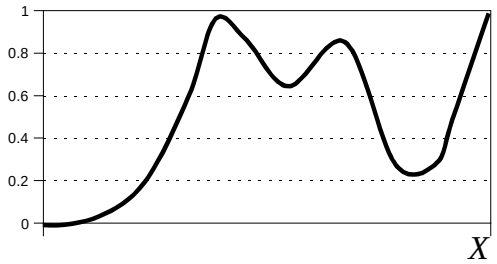
Knowledge Context







b

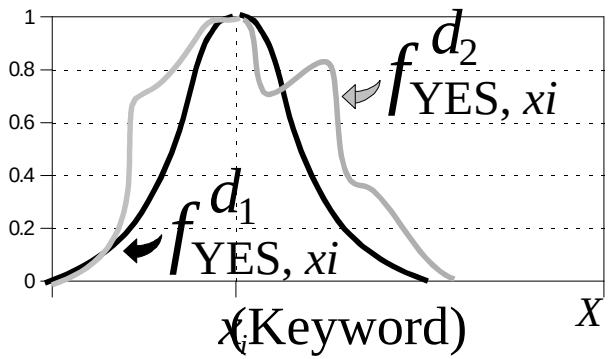


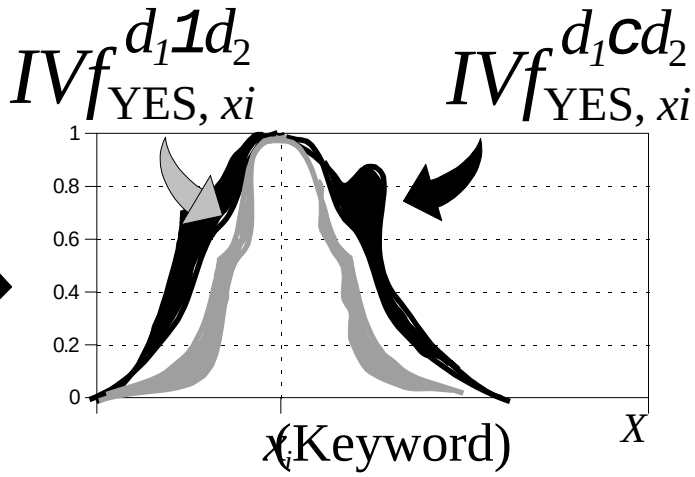
Single Knowledge Context $+\mathbf{X}, d_i$

2 Information Resources

With Evidence Sets

2 Local Knowledge Contexts with semi-metrics d_1 and d_2 with equal weight.





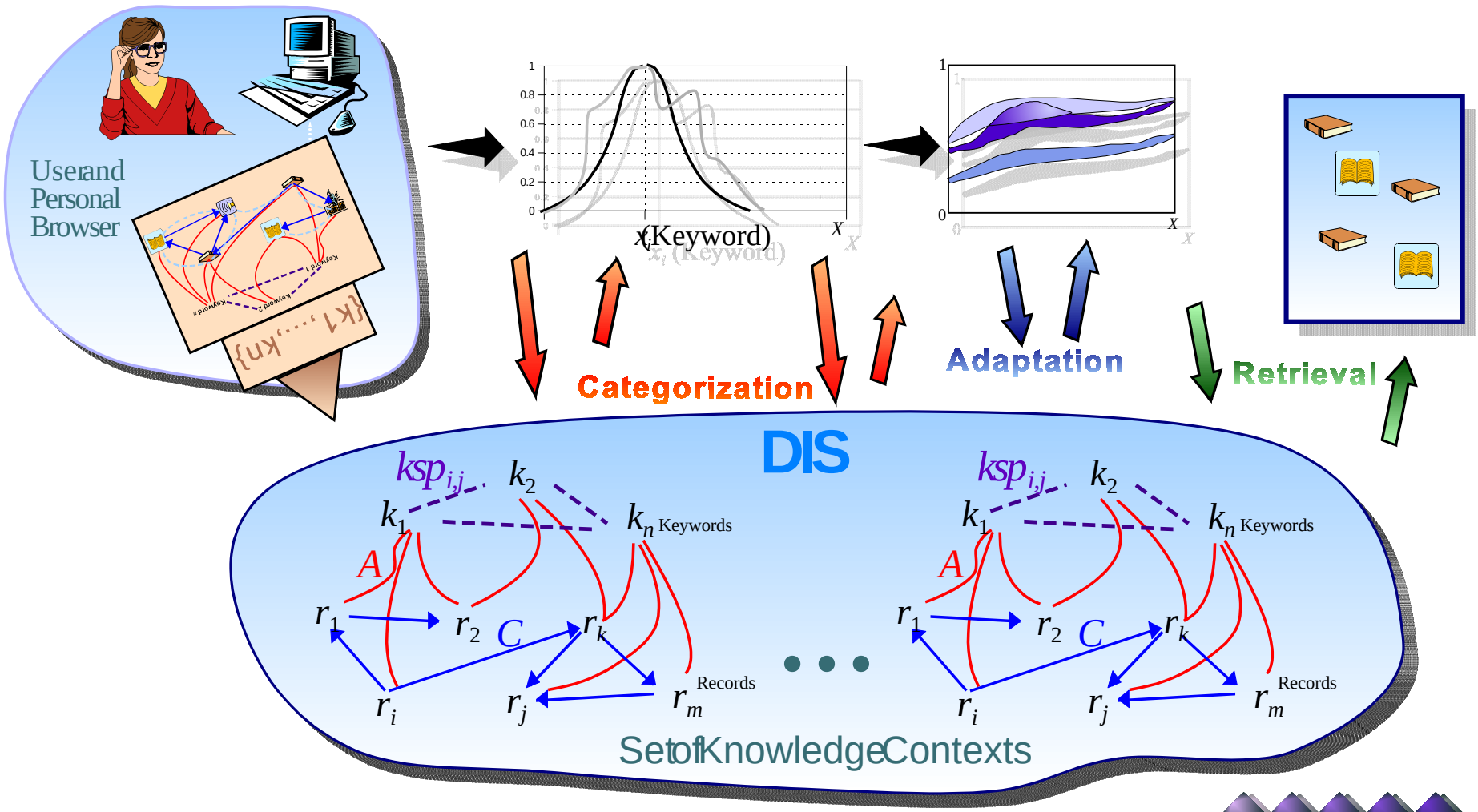
Intermediate Fuzzy Sets Associated with each Context

YES Evidence Set: The weights of the basic probability assignment are 0.5 for both intervals $(m_1+m_2)/(2*1)$

Integration of two local knowledge contexts: information resource and agent

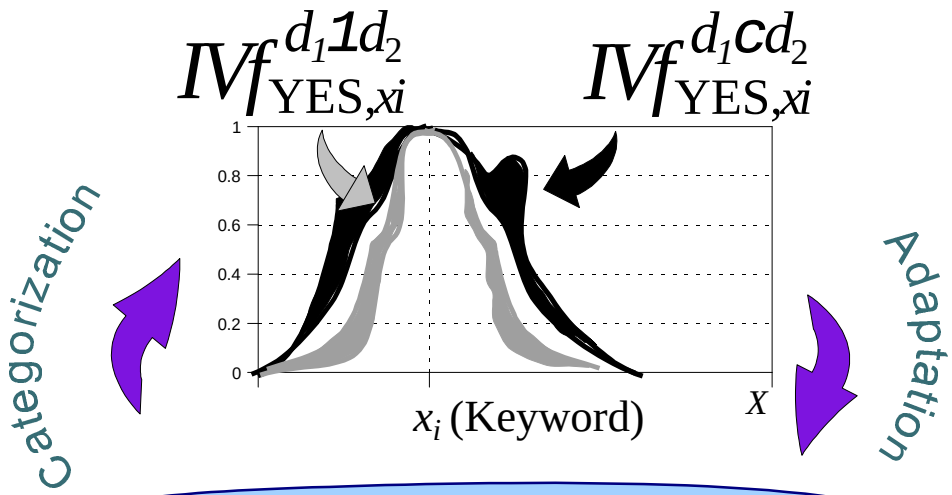
TalkMine: Agent-DIS Conversation

Usage and Adaptation of Keyword Semantics



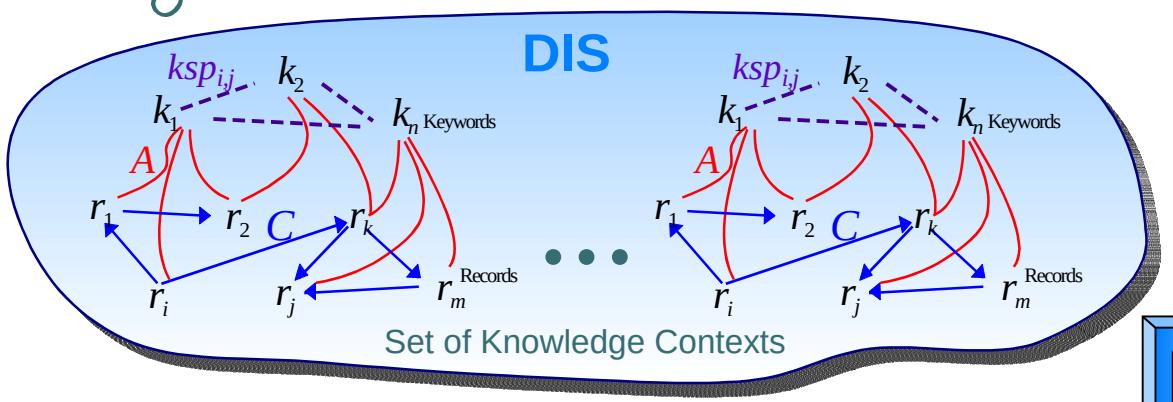
Semantic Adaptation

Short-Term Categorization



Long-Term Memory

New keywords are recognized by Information Resources

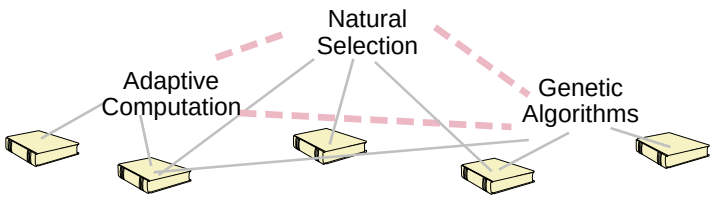


$$N^k_{t\%1}(x_i) \leftarrow N^k_t(x_i) \% w.A(x_i),\ k \leftarrow 1...n_d,\ x_i \in X$$

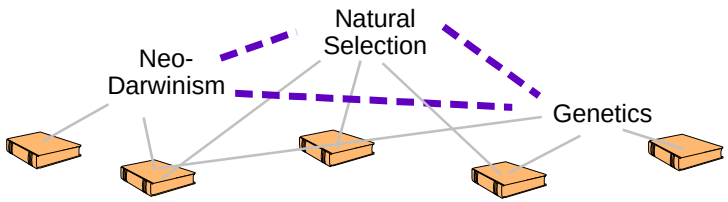
$$N^k_{t\%1}(x_i,x_j) \leftarrow N^k_t(x_i,x_j) \% w.\min[A(x_i),A(x_j)],\ k \leftarrow 1...n_d,\ x_i \in X$$

Evolving Knowledge

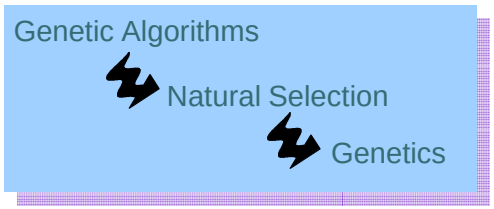
Categorical Representation

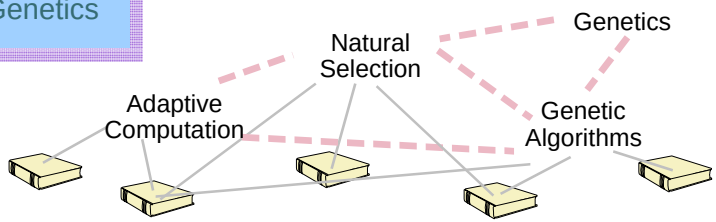


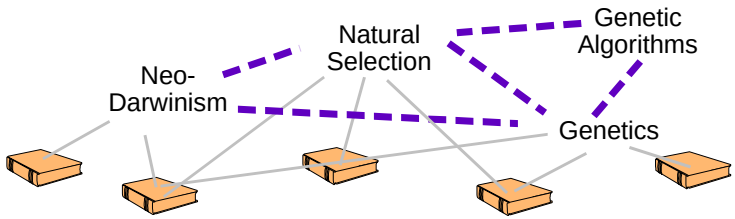
SFI



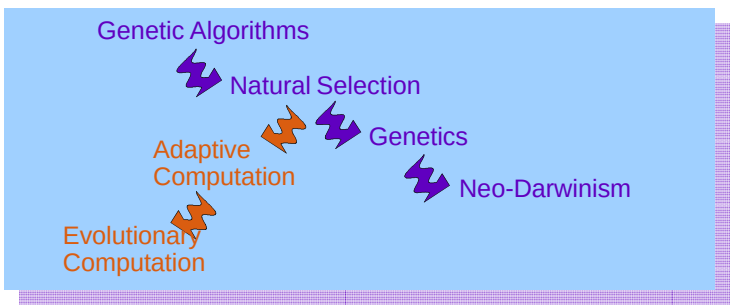
Philosophy of Biology







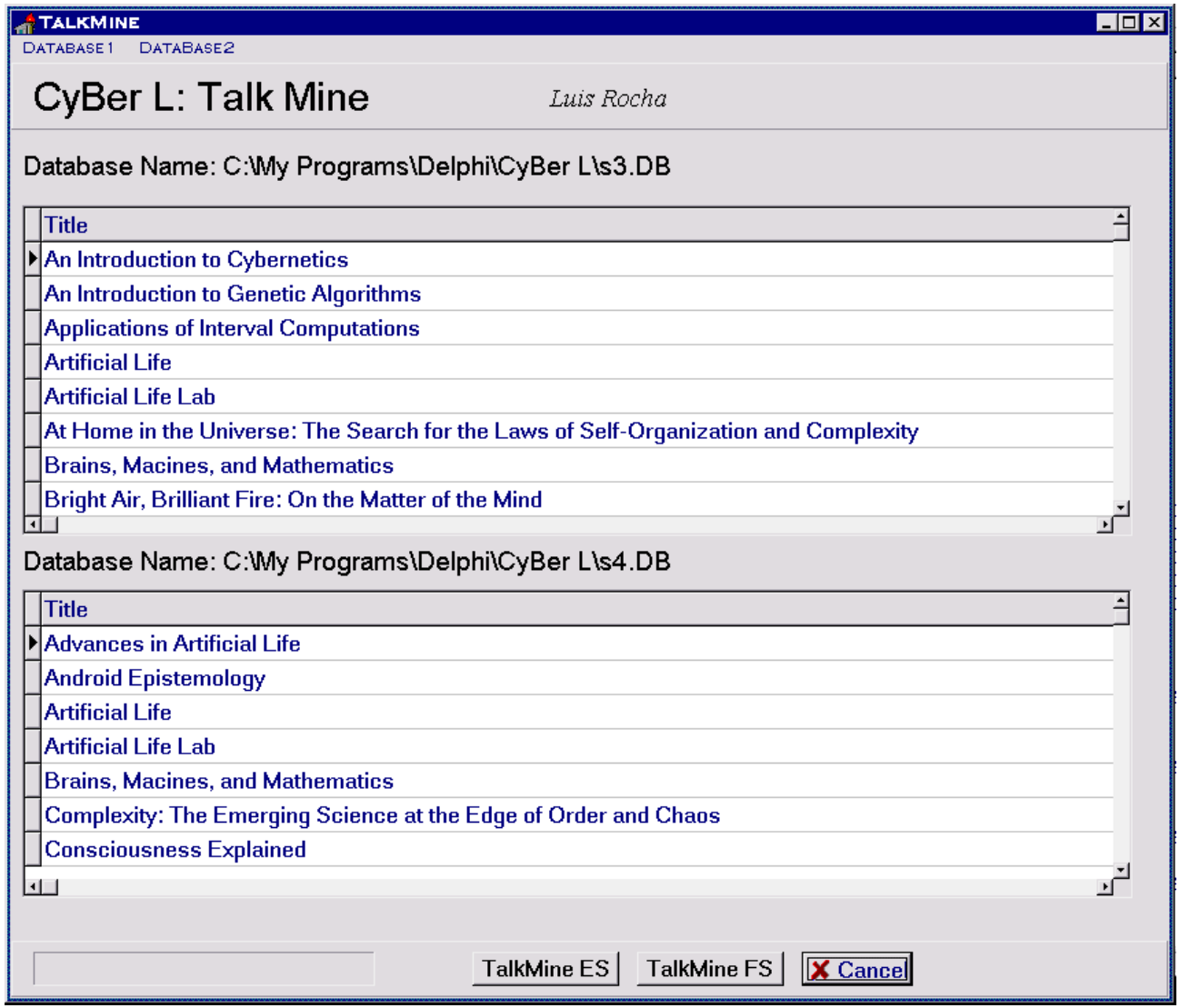
Information Crossover (Creativity)



Network Association

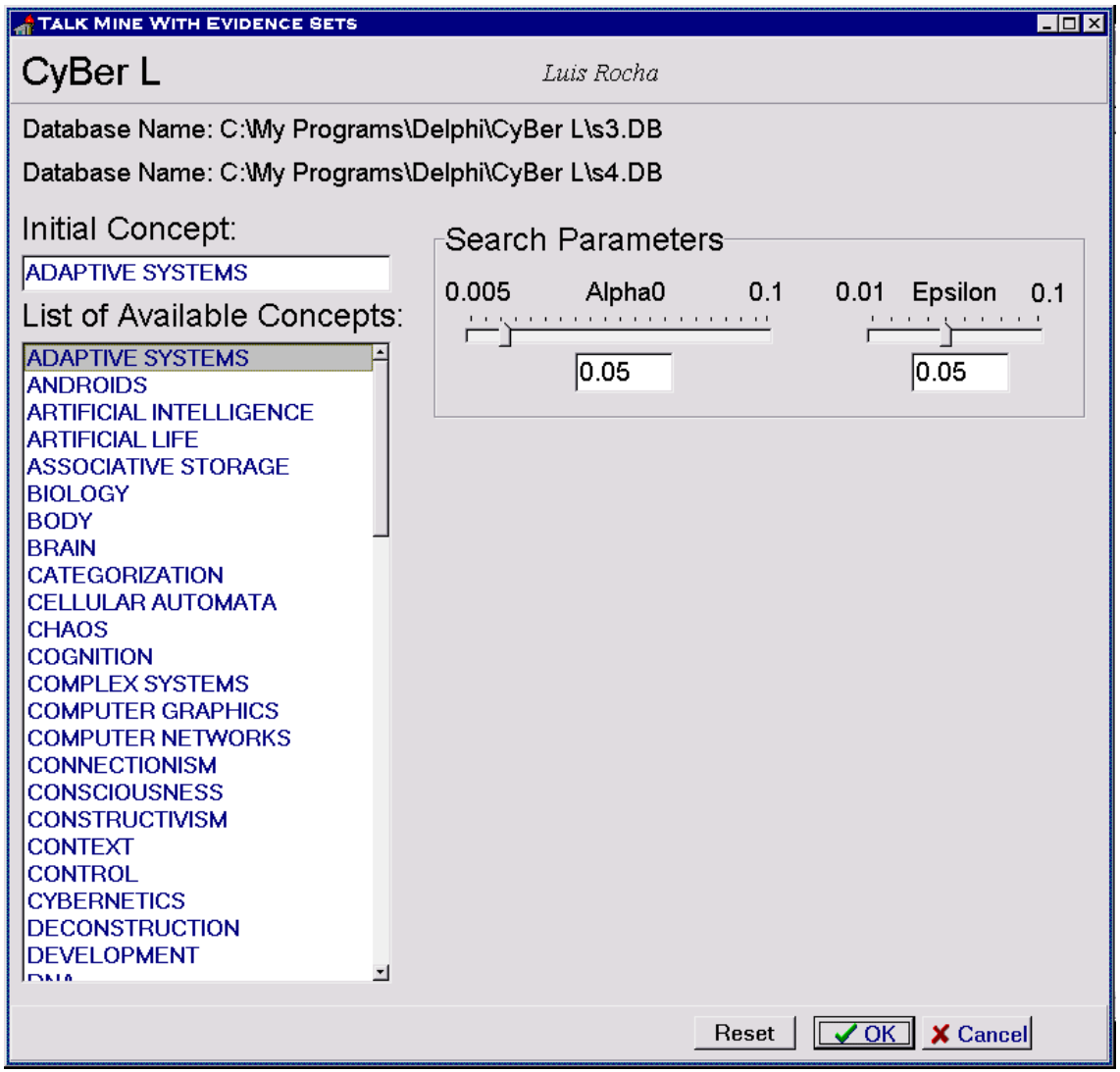
TalkMine

WithTwoLibraryDataBases



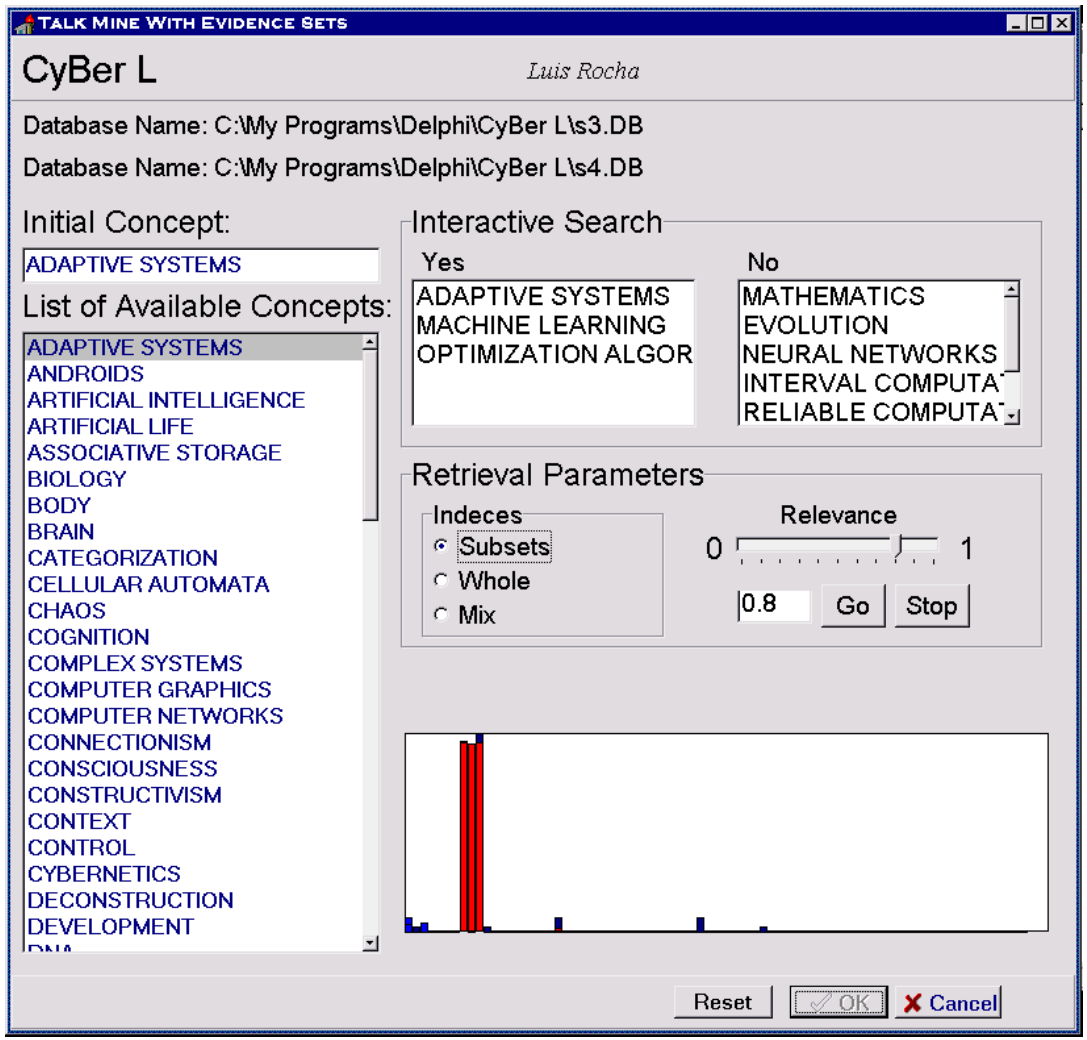
TalkMine

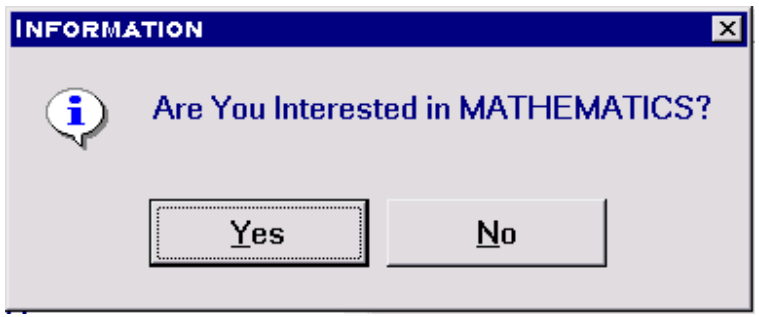
Search Screen



TalkMine

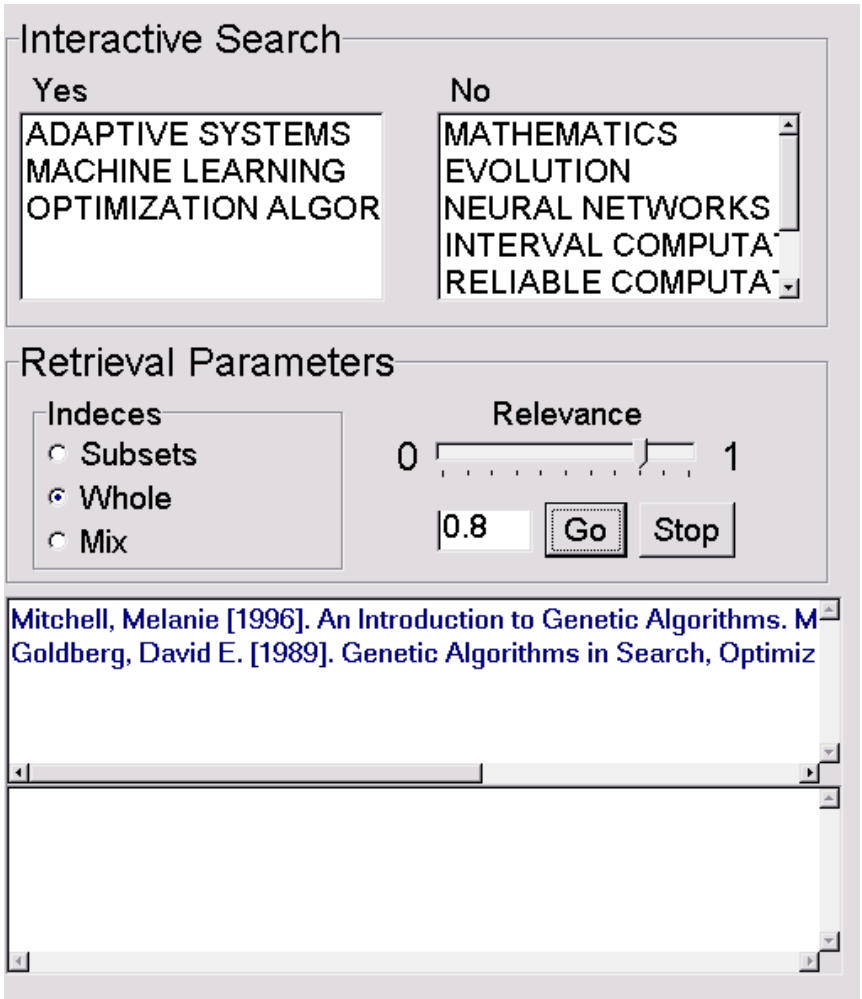
Question and Retrieval Screen

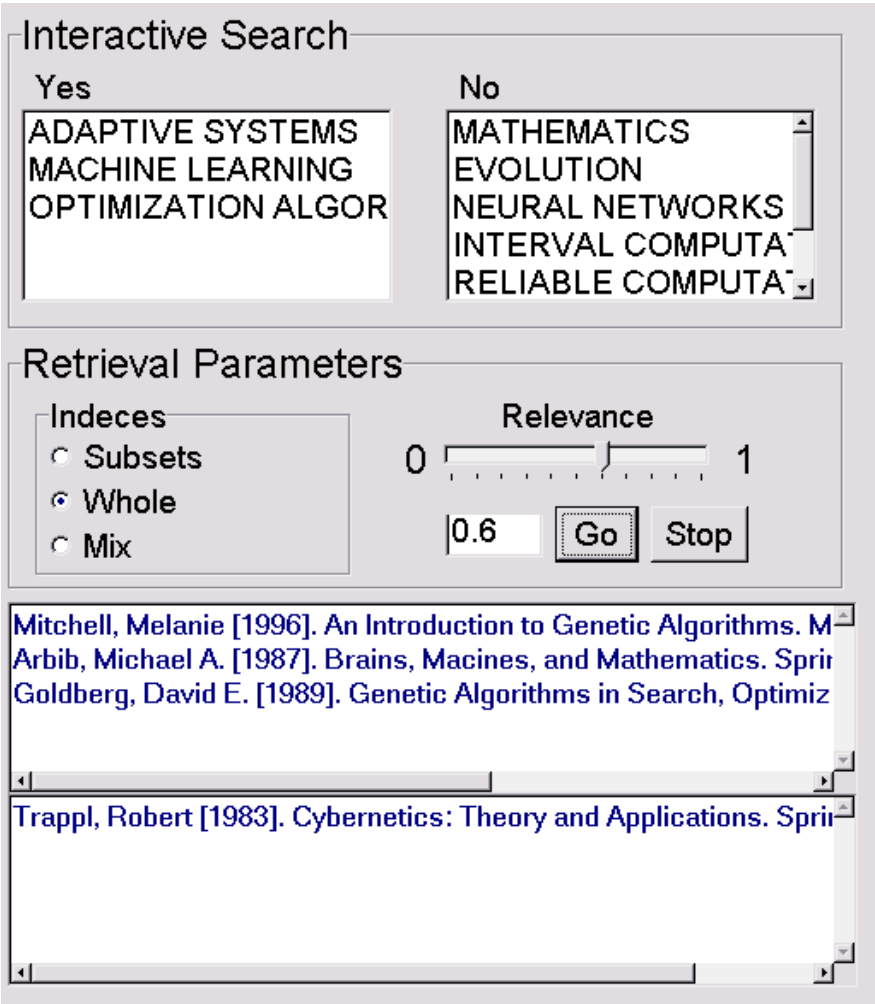




TalkMine

Result Screens

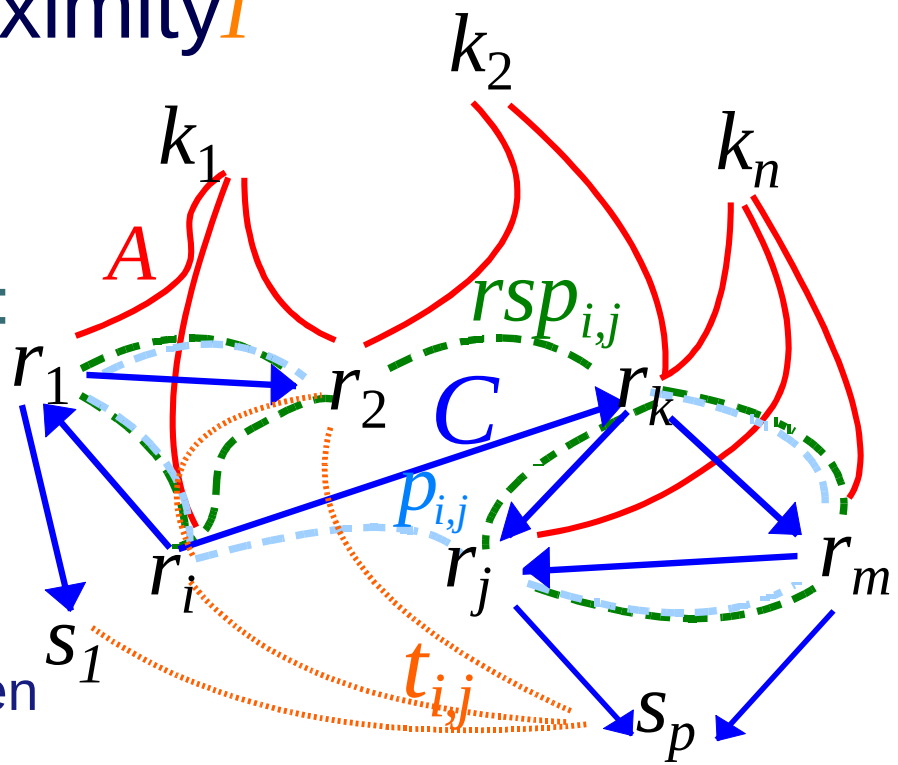




Structural Adaptation via Collective Interaction

Calculate User Traversal Proximity T

- Obtain n user paths of 3 documents retrieved sequentially by the same user
- For each path apply 3 *learning rules*:
  - *Frequency*: Proximity between two sequentially retrieved documents is reinforced
  - *Symmetry*: When proximity between 2 documents is reinforced, the symmetrical direction is partially reinforced
  - *Transitivity*: If the proximity values between d1 and d2 and d2 and d3 increase, the proximity between d1 and d3 also increases (less).



Recommendations with Spreading Activation

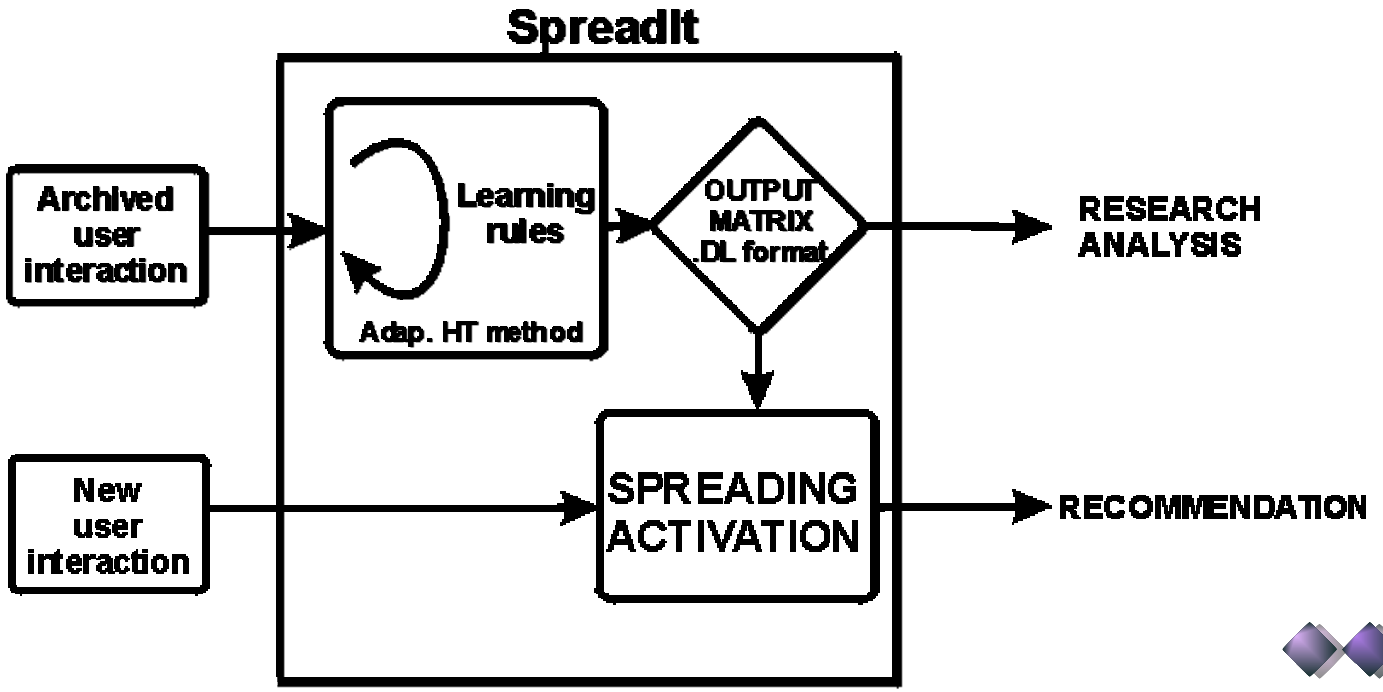
On Proximity Networks

- Selected words in the network get an activation value which follows the links to associated words. The activation reaching a word is the sum of the weighted activations coming through all its input links. From an activated word, activation diffuses further to the words linked to it, and so on.
  - < Anderson (1973, 1984), (Klimesch, 1994)
- Finds nodes related to all the nodes that are initially activated, according to a particular proximity network organization.
  - < Essentially, constructs a clique of related nodes using linear algebra.
- Much more efficient (and more brain-like) than the traditional keyword match pursued by search engines.
  - < Provided we have good proximity networks

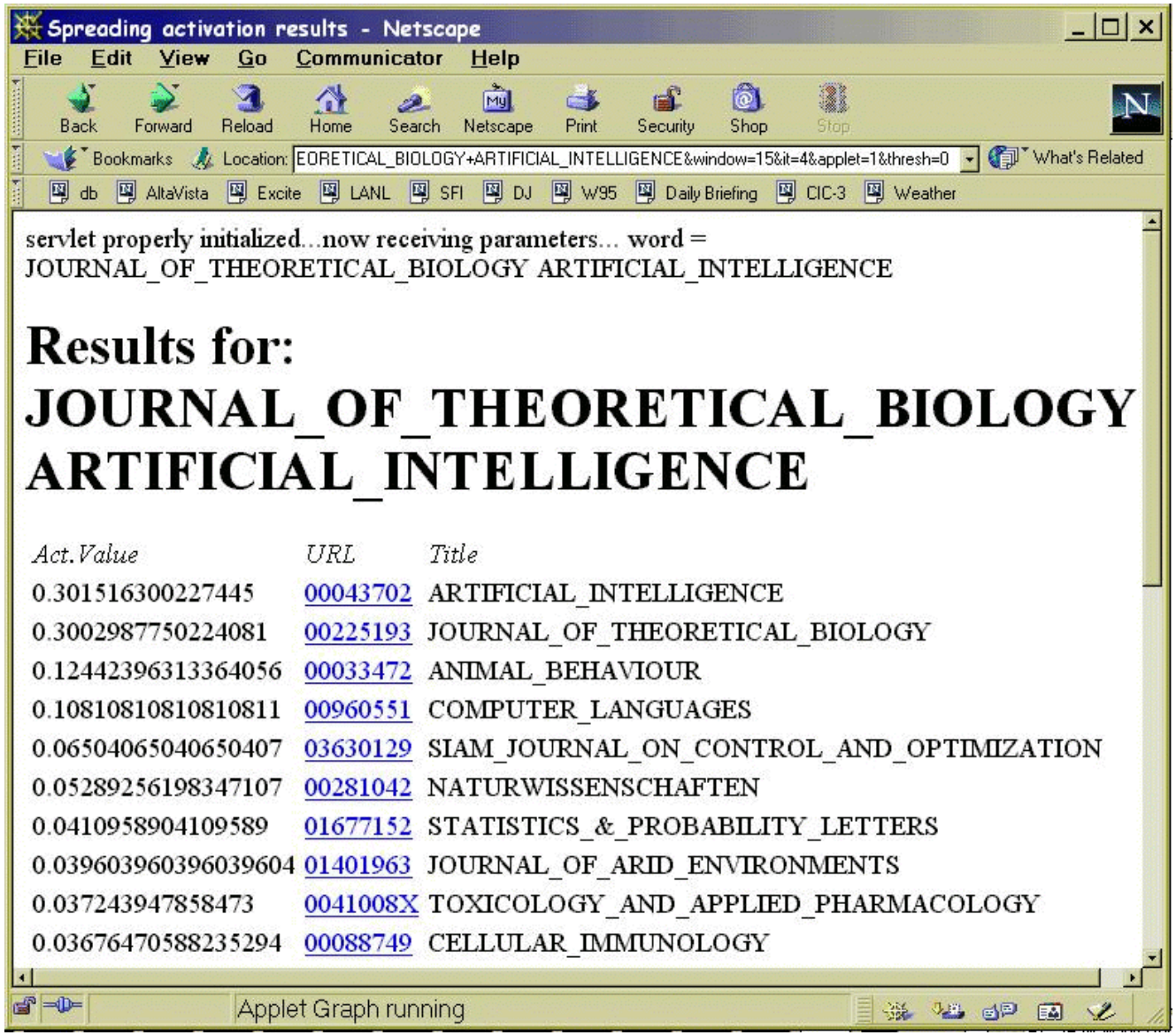
The ISSN Experiments

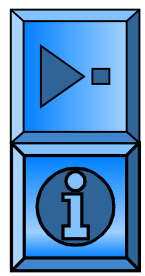
@ApWeb and Spreading Activation

- ISSN Journals Database (Approx 800)
  - < Build Network from daily user records using *@ApWeb*
  - < Journals searched by same user on same day



SANewInterface





SARResults

ThePrincipiaCyberneticaWeb(cont)

matchingkeytermsfor"beginlife"
variabledefinedStartspreading...
donespreading!...
showingresults...
0.45454545454545455SEMANASemantic_Analysis
0.44166666666666665COGNALLinks_on_Cognitive_Science_and_AI
0.25833333333333336MULTAXIMultiple_axiomatization_sets,_a_metaphor_for
0.25714285714285712GODEXISArguments_for_and_against_the_Existence_of_God
0.22222222222222224METHODOMethodology_for_the_Development_of_MSTT
0.21428571428571427FUTDEVLinks_on_Future_Development
0.17970177643707058MEAN1STMeaning_Goes_First_[empty]
0.16666666666666666MULTICELMulticellular_organisms
0.16363636363636364MSTQUANMST_as_the_quantum_of_evolution_[empty]
0.16320346320346318CONBUILConsensus_Building
0.15961954108330412ORILIFETheOrigins_of_Life
0.15625SEXSexuality_as_a_Metasystem_Transition
0.13333333333333333BIOEVOLBiological_Evolution
0.13043478260869565HISTORYHistory_of_the_Principia_Cybernetica_Project
0.12961750241038013FOUNDCONFoundational_Concep

SAPCP

Spreading activation results - Netscape

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Shop Stop

Bookmarks Location: ctKeyterms_servlet3?word=Artificial+Complex_Adaptive_Systems&window=15&it=4&applet=1 What's Related

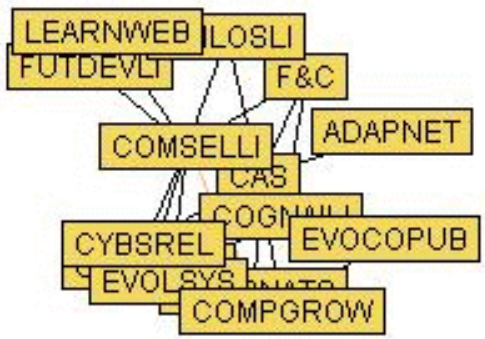
db AltaVista Excite LANL SFI DJ W95 Daily Briefing CIC-3 Weather

servlet properly initialized...now receiving parameters... word = Artificial Complex_Adaptive_Systems

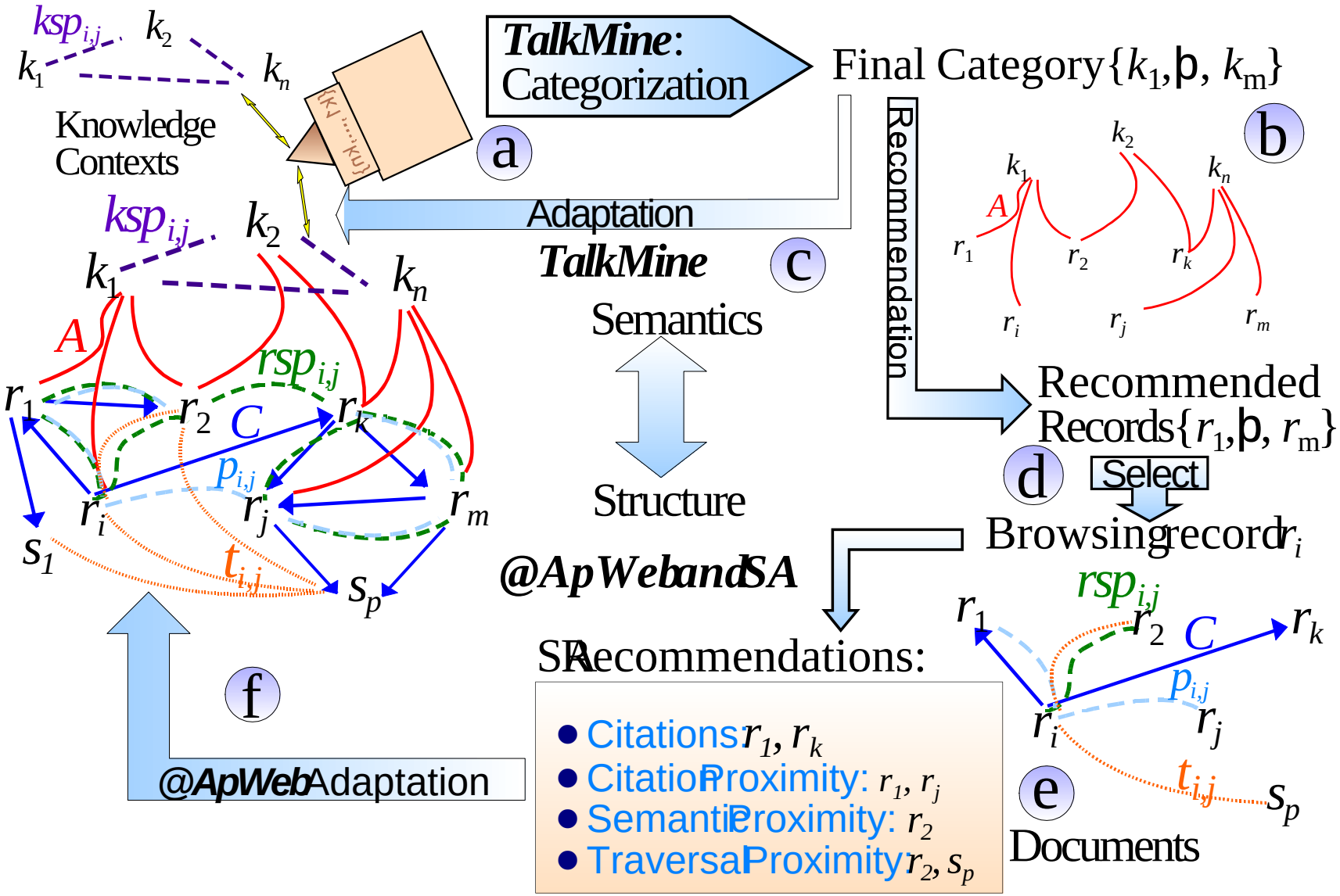
Results for: Artificial Complex_Adaptive_Systems

Act.Value	*URL*	*Title*
0.5545454545454545	COMPNATS	Self-organization_and_complexity_in_the_natural_sciences
0.44166666666666665	COGNAILI	Links_on_Cognitive_Science_and_AI
0.3	ALEVOMOD	Artificial_Life_Evolutionary_Models
0.3	CAS	Complex_Adaptive_Systems
0.3	COMSELLI	Links_on_Complexity,_Self-organization_and_Artificial_Life
0.225	COMPLEXI	What_is_complexity?
0.21428571428571427	ADAPNET	Adaptive_hypertext_network
0.21428571428571427	F&C	Form_and_Content
0.21428571428571427	FUTDEVLI	Links_on_Future_Development
0.11764705882352941	EVOLSYS	Evolutionary_Systems
0.10526315789473684	PHILOSLI	Links_on_Philosophy
0.0868421052631579	CYBSREL	Relation_to_other_disciplines
0.07575757575757576	EVOCOPUB	Classic_Publications_on_Complex,_Evolving_Systems

Applet Graph running



Adaptation of Structure and Semantics using the Collective Behavior of Users



Distributed Active Recommendation

TalkMine and @ApWeb provide Collective Organization of DIS with Enabling Relations

- ***Active environment*** of user-system interaction capable of identifying users/agents and recommending relevant information.
- Explores ***Structural relationships*** with proximity measures, adaptive via *@ApWeb*.
- It establishes an ***evolving semantics*** as keyword associations adapt to the expectations of users and new keywords are exchanged among multiple information resources and users browsers with *TalkMine*.
- It establishes ***linked information resources*** as users search several resources simultaneously and establish all-way information exchanges.

Beyond Information Retrieval

Biologically Motivated Design

- ***Recommendation***: the system pro-actively pushes relevant documents to users about related topics that they may have been unaware of.
- ***Conversation*** between users and information resources and among information resources (and indirectly among users) with interactive categorization.
- ***Creativity***. New semantic and structural associations are set up by *TalkMine* and *@ApWeb*. The categorization process brings together knowledge from the different information resources. This not only adapts existing knowledge, but combines knowledge not locally available to individual information resources. In this sense, because of the conversation process, information resources gain new knowledge previously unavailable.